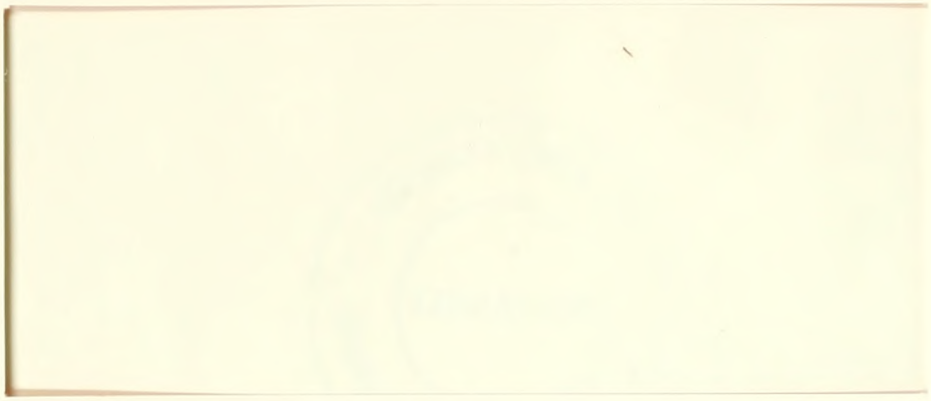
WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

ASYMPTOTIC PROPERTIES OF BIVARIATE K-MEANS CLUSTERS

M. Anthony Wong

W.P. #1216-81 May 1981

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139



ASYMPTOTIC PROPERTIES OF BIVARIATE K-MEANS CLUSTERS

M. Anthony Wong

W.P. #1216-81 May 1981

ASYMPTOTIC PROPERTIES OF BIVARIATE K-MEANS CLUSTERS

M. Anthony Wong

Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

Key Words and Phrases: k-means clusters; within cluster sum of squares; graph theory; regular hexagons.

ABSTRACT

A bounded region in R^2 with a uniform density function defined over it is partitioned into k sub-regions such that the within cluster sum of squares is minimized. An asymptotic ($k \to \infty$) lower bound for the within cluster sum of squares of this optimal k-means partition is obtained. This lower bound is useful in suggesting that the graph-configuration of the optimal k-partition would consist of regular hexagons of equal size when k is large enough. An empirical study illustrating these asymptotic properties of bivariate k-means clusters is also presented.

1. INTRODUCTION

Let the observations $x_1,\ldots,x_N$ be sampled from a distribution F with density function f. In cluster analysis, the k-means clustering method (see Hartigan (1975), Chapter 4) is often used to partition the sample of N observations into k clusters with means $\bar{x}_1,\ldots,\bar{x}_k$. The resultant clusters satisfy the property that no movement of an observation from one cluster to another reduces the sample within cluster sum of squares

$$WSS_N = \sum_{i=1}^{N} \min_{1\le j\le k} || x_i - \bar{x}_j||^2/(N-k)$$

For these k-means clusters, a k-partition of the sampled space can be defined by associating each cluster mean $\bar{x}_j$ with the convex polyhedron C_j of all points in R^p closer to x_j than to any other cluster mean. The corresponding optimal k-means partition in the population F is defined by the population cluster means $\mu_1,\ldots,\mu_k$, which are selected in such a way that the within cluster sum of squares

$$WSS = \int \inf_{1\le j\le k}||x - \mu_j||^2 \, dF$$

For fixed number of clusters k, the asymptotic convergence (as $N \to \infty$) of the sample k-means clusters to the population k-means clusters has been studied by MacQueen (1967), Hartigan (1978), and Pollard (1981). The most recent result can be found in Pollard (1981), in which conditions are found that ensure the almost sure convergence of the set of means of the k-means clusters. However, the asymptotic properties of k-means clustering in the case where the number of cluster k increases with the sample size N did not receive much attention.

In Hartigan and Wong (1979a) and Wong (1980), some asymptotic properties (as $k \to \infty$) of the population k-means clusters in one dimension are obtained. It is shown that the within sum of squares of the k clusters are asymptotically

equal, and that the length of the jth cluster interval ($1 \le j \le k$) is inversely proportional to $f(c_j)^{1/3}$, where c_j is the midpoint of the jth cluster interval. It is also shown that if $k(N) \to \infty$ as $N \to \infty$ with $k(N) = o(N/\log N)^{1/3}$, then the sample k-means clusters have asymptotic properties similar to that of the population clusters. Using these results, it can also be shown that a uniformly consistent histogram estimate of f, which is constant over each k-means cluster interval, can be constructed from the sample using the k-means method.

Unfortunately, these univariate results cannot be easily generalized to the multivariate case. Only empirical evidence exists to support the conjecture that similar asymptotic results hold in several dimensions for k-means clusters, and that a uniformly consistent histogram estimate of a multivariate density f can be constructed by the k-means method. The latter uniform consistency result is of special practical interest as it would justify the usage of the computationally efficient k-means method (Hartigan and Wong, 1979b) for estimating multivariate density functions from large samples.

In this paper, some asymptotic properties of the population k-means clusters for uniform distributions in R^2 are given. In Section 2, an asymptotic lower bound for the WSS of the optimal k-means partition is obtained. Since this lower bound is attained when all k clusters of the partition are regular hexagons of equal area, this result suggests that the graph configuration of the optimal k-means partition would consist of regular hexagons when k is large enough. An empirical study is performed to illustrate these asymptotic properties of bivariate k-means clusters, and the results are given in Section 3.

The results given in this paper fall short in generalizing the asymptotic properties of the univariate population k-means clusters to the bivariate case. However, they are the first results obtained in the investigation of the properties of bivariate k-means clusters.

2. AN ASYMPTOTIC LOWER BOUND FOR THE WITHIN CLUSTER SUM OF SQUARES OF K-MEANS CLUSTERS

In this paper, some asymptotic properties $(k \to \infty)$ of the population k-means clusters for the uniform density in two dimensions are given. The main result is Theorem 2, which gives an asymptotic lower bound for the within cluster sum of squares of the optimal k-means partition.

Theorem 2:

Let $\mathcal{A}$ be a region of area A with a connected interior in R^2. Suppose that the boundary of A has finite length and let $1/A$ be the constant density over $\mathcal{A}$.

Let WSS be the minimum within cluster sum of squares over all k-partitions. Then

$$\lim_{k\to\infty} WSS/(\frac{5\sqrt{3}}{54} \cdot \frac{A}{k}) = 1 .$$

(Remark: Since the asymptotic lower bound given in Theorem 2 is attained when all k clusters of the partition are regular hexagons of equal area, this result suggests that the graph configuration of the optimal k-means partition would consist of regular hexagons when k is large enough.)

In outline, the proof of Theorem 2 requires first showing that the polygon with n edges and area A which has the minimum within polygon sum of squares is regular (see Lemma 1). For any polygon divided into k clusters, a lower bound for the limiting value (lim inf) of the within cluster sum of squares may then be found by assuming the clusters are regular hexagons (Theorem 1). An upper bound may also be found by covering the polygon with regular hexagons. Since the ratio of the two bounds approaches 1 as $k \to \infty$, Theorem 2 follows.

Hence, to show the result in Theorem 2, we need the following lemmas:

LEMMA 1

If f is the constant density over an n-sided polygon $\mathcal{A}$ of area A, then a lower bound for $WPSS_{\mathcal{A}}$, the within polygon sum of squares of $\mathcal{A}$, is given by $\frac{1}{2n} fA^2[\frac{1}{\tan \pi/n} + \frac{1}{3} \tan \frac{\pi}{n}]$.

However, to prove Lemma 1, we need Lemma 1.1 and Lemma 1.2.

Lemma 1.1

For a triangle Δ, with fixed area A_o, and fixed angle θ_o at the vertex V, the minimum value of $\int_\Delta r^2 d(\text{Area})$ (where r is the distance from V) is $\frac{1}{2}A_o^2[\frac{1}{\tan 1/2\ \theta_o} + \frac{1}{3} \tan \frac{1}{2} \theta_o]$, achieved when Δ is isosceles with equal edges adjacent to V.

Proof.

Consider the triangle VTS with angle $SVT = \theta_o$ and has an area of A_o. Let M be the midpoint between the vertices T and S (see Fig. 1). Without loss of generality, let M be at the origin. Let $\underline{e}$ be the unit vector along the base ST (x-axis) and put $\|TM\| = t$. Also, let V be represented by $\underline{z}$. It follows that if ST is rotated by $d\theta$ about M to S'T' (see Fig. 1), the

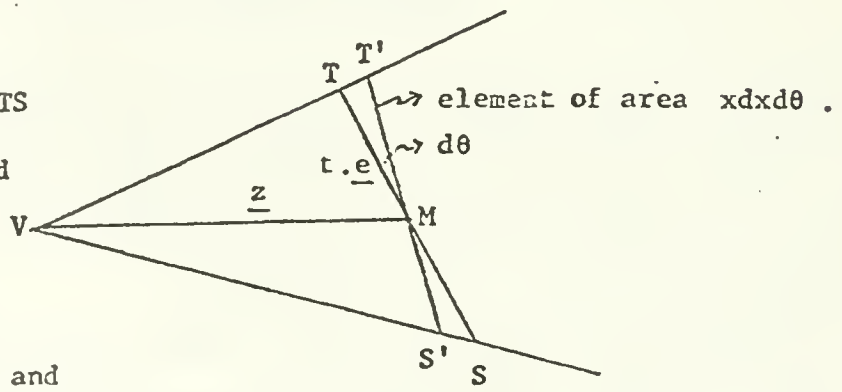


Fig. 1

increment in $\int r^2 d$ (Area), the second moment about V, is given by

$$\int_o^t \|\underline{z} + x\underline{e}\|^2 \, x dx d\theta - \int_o^t \|z - x\underline{e}\|^2 \, x dx d\theta$$

$$= \int_o^t 4(\underline{z}\cdot\underline{e}) \, x^2 dx d\theta$$

$$= \frac{4}{3} (\underline{z}\cdot\underline{e}) t^3 \quad .$$

Thus, this increment is positive or negative as $\underline{z}\cdot\underline{e}$ is positive or negative. And hence the minimum occurs when $\underline{z}\cdot\underline{e} = 0$; that is, when the triangle is isosceles.

Now, if we move towards the isosceles position by such a rotation, the area increases. Thus, for a given triangle VTS, we can decrease the 2nd moment about V by first rotating to the isosceles position, and then sliding the base back towards V until the triangle has area A_o. Since the 2nd moment about V for an isosceles triangle of area A_o is equal to $\frac{1}{2}A_o^2 \left[\frac{1}{\tan 1/2\ \theta_o} + \frac{1}{3} \tan \frac{1}{2}\theta_o \right]$, the lemma follows.

Lemma 1.2

Let VT and VS be two lines in $\mathbb{R}^2$ with angle TVS = $\theta_o < \Pi$. Suppose that Q is a union of quadrilaterals (whose interiors are disjoint) all of whose vertices lie on VT and VS. Let the area of Q be A_o. Then $\int_Q r^2 d$ (Area), the second moment about V, is minimized when Q is an isoceles triangle.

Proof.

[I] Fix an integer i and real number u, and let $\mathcal{Q}(i,u)$ be the set of plane figures such that every $Q \in \mathcal{Q}(i,u)$ is a union of quadrilaterals (whose interiors are disjoint), all of whose vertices lie on VL_1 or VL_2; L_1 and L_2 lie respectively on VT and VS, with $\|VL_1\| = \|VL_2\| = u$. Using the labeling system shown in Fig. 2, it

is clear that every $Q \in \mathcal{Q}(i,u)$ is uniquely determined by the set of vertices $(x_1, \ldots, x_{4i})$, where the x_j's satisfy

(i) $0 \le x_1 \le \cdots \le x_{2i} \le u$,

and

(ii) $0 \le x_{2i+1} \le \cdots \le x_{4i} \le u$.

Thus, $\mathcal{Q}(i,u)$ can be identified with a compact subset of $[0,u]^{4i}$, from which it inherits a natural topology (pointwise convergence of the x_j's). Under this topology, the two mappings $f_a : \mathcal{Q}(i,u) \to \mathbb{R}$ and $f_w : \mathcal{Q}(i,u) \to \mathbb{R}$, where $f_a(Q)$ = area of Q and $f_w(Q) = \int_Q r^2 d(\text{Area})$, are continuous on $\mathcal{Q}(i,u)$. Therefore, if $A_o \le$ area of triangle VL_1L_2, the set

$$\mathcal{Q}_o(i,u) = \{Q \in \mathcal{Q}(i,u) \text{ such that } f_a(Q) = A_o\}$$ is nonempty and compact.

It follows that $\min_{Q \in \mathcal{Q}_o(i,u)} f_w(Q)$ is attained by some $Q_o \in \mathcal{Q}_o(i,u)$.

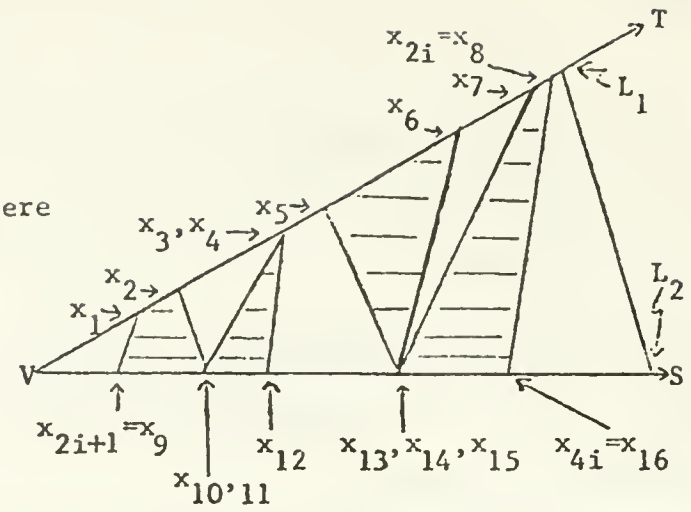


Fig. 2 An Example of an $Q \in \mathcal{Q}(4,u)$. (x_j's are the distances from V.)

[II] Next, we will show that, for any i and u, Q_o is an isosceles triangle. It is enough to show that the result holds when $i = 2$. Suppose that Q_o were not an isosceles triangle. Using Lemma 1.1 to eliminate other cases, it is sufficient to consider the case when Q_o = triangle VBA $\cup$ triangle AFH (see Fig. 3), where

(1) $\|VB\| \ge \|VA\|$, and

(2) $\|VF\| \ge \|VH\|$.

Consider the triangle AFH . Choose a point C on BF , close to F . Let the line through C parallel to VS cut the triangle AFH along the segment C*G* . Translate C*G* along this line to form a trapezium CGHA with $\|CG\| = \|C^*G^*\|$. By this construction (see Fig. 3), the trapeziums CGHA and C*G*HA have the same area, but $\int_{CGHA} r^2 \, d(\text{Area}) < \int_{C^*G^*HA} r^2 d(\text{Area})$.

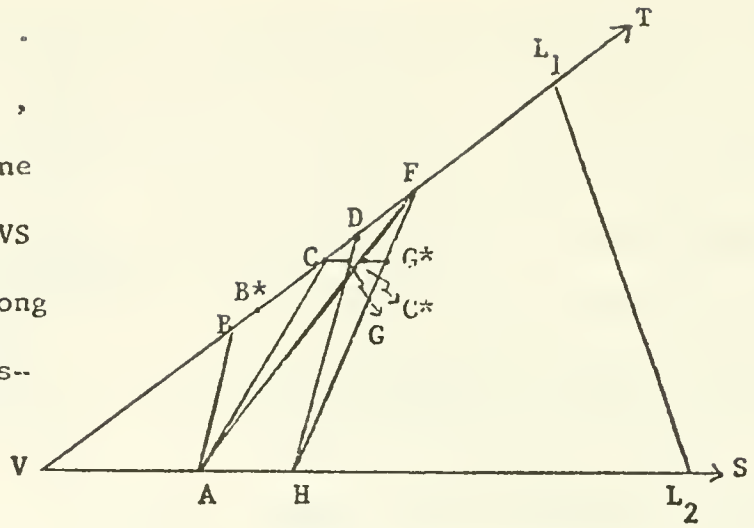


Fig. 3

Produce HG to intersect VT at D . Since $\|CG\| = \|C^*G^*\|$, triangle C*FG* has a larger area than triangle CDG . Part of the area of C*FG* can therefore be redistributed to complete the quadrilateral CDHA with a decrease in 2nd moment. The remaining area can be added to triangle VBA to produce triangle VB*A . If $\|CF\|$ is small enough, all the points of triangle C*FG* are further from V than all the points of triangle ABB* ; 2nd moment is thus decreased.

[III] From the result of [I] and [II], $\min_{Q \in \mathcal{Q}_o(i,u)} f_w(Q)$ is attained by Q_o , the isosceles triangle with equal edges adjacent to V . Notice that it is the same Q_o for each $\mathcal{Q}(i,u)$. Thus Q_o gives the minimum value of $\int r^2 d(\text{Area})$ over $\bigcup_{i,u} \mathcal{Q}(i,u)$.

Now, we can proceed to obtain the result given in Lemma 1.

Proof of Lemma 1.

[I] Consider a given n-sided polygon $\mathcal{A}$ of area A. By joining the centroid C and each of the n vertices of $\mathcal{A}$, we obtain an n-partition of $\mathcal{A}$ defined by the n cones radiating from C (see Fig. 4). Let $\mathcal{A}_i$ be the subset of $\mathcal{A}$ in the ith cone. Let A_i be the area of $\mathcal{A}_i$ and let $2\theta_i$ $(<\Pi)$ be the angle subtended at C by the ith cone. Then

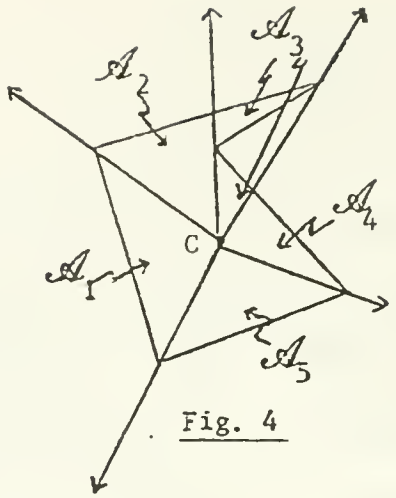


Fig. 4

$\sum_1^n A_i = A$ and $\sum_1^n \theta_i = \Pi$. From Lemma 1.1 and Lemma 1.2 we have for each $1 \le i \le n$,

$$\int_{\mathcal{A}_i} r^2 \, d\,(\text{Area}) \ge \frac{1}{2} A_i^2 \left[\frac{1}{\tan \theta_i} + \frac{1}{3} \tan \theta_i\right],$$

where r is the distance from C. Summing over i, we have

$$(1)\quad \text{WPSS}_{\mathcal{A}} = f \sum_1^n \int_{\mathcal{A}_i} r^2 d\,(\text{Area}) \ge \frac{1}{2} f \sum_1^n A_i^2 \left[\frac{1}{\tan \theta_i} + \frac{1}{3} \tan \theta_i\right].$$

[II] Next, we will find the minimum of $\sum_1^n A_i^2 \left[\frac{1}{\tan \theta_i} + \frac{1}{3} \tan \theta_i\right]$ under the constraints: (i) $\sum_1^n A_i = A$ and (ii) $\sum_1^n \theta_i = \Pi$; $\theta_i < \frac{\Pi}{2}$ for $1 \le i \le n$. Now, using Lagrange multipliers, the minimum must satisfy:

$$(2)\quad A_i \left[\frac{1}{\tan \theta_i} + \frac{1}{3} \tan \theta_i\right] = c_1 \text{ ; and}$$

$$(3)\quad A_i^2 \left[-\frac{\sec^2 \theta_i}{\tan^2 \theta_i} + \frac{1}{3} \sec^2 \theta_i\right] = c_2,$$

where c_1 and c_2 are constants independent of i.

Thus, squaring (2) and then dividing by (3), we have

$$(1 + \frac{1}{3}\tan^2\theta_i)^2 \tan^{-2}\theta_i / [-(1 + \tan^{-2}\theta_i) + \frac{1}{3}(1 + \tan^2\theta_i)]$$

$$= C_1^2/C_2 = C_3 ,$$

and hence

$$(4) \qquad (\frac{4}{9}\tan^4\theta_i)/(1 + \frac{2}{3}\tan^2\theta_i + \frac{1}{9}\tan^4\theta_i) = 1 + \frac{1}{C_3} = C_4 .$$

Since the left side of (4) is a strictly increasing function of $\tan^2\theta_i$, $\tan^2\theta_i$ is a constant.

But $\theta_i < \pi/2$ for all i, so we must have $\theta_i = \pi/n$ for all $1 \le i \le n$.

Also, $A_i = A/n$ for all $1 \le i \le n$.

[III] Using the result of [II], we have from (1) that

$$WPSS_{\mathcal{A}} \ge \frac{1}{2} f \sum_1^n A_i^2 [\frac{1}{\tan\theta_i} + \frac{1}{3}\tan\theta_i]$$

$$\ge \frac{1}{2n} f A^2 [\frac{1}{\tan \pi/n} + \frac{1}{3}\tan\frac{\pi}{n}] ,$$

and the equality holds when the given n-polygon $\mathcal{A}$ is regular, which gives the lemma.

(Remark: In the application of this lemma, f is usually the constant density over a region containing the n-sided polygon $\mathcal{A}$. Thus, $f < 1/A$ for most applications.)

Next, in Theorem 1, we will obtain a lower bound for the within cluster sum of squares of the optimal k-means partition of a polygon in R^2. However, it is important to first establish that it is sufficient to consider only "3-edge" k-partitions (k-partition whose corresponding graph configurations have the property that all the interior vertices are associated with exactly 3 edges.) Hence, we need Lemma 2.

Lemma 2

Let $\mathcal{A}$ be a region with connected interior in R. Suppose that the constant density over $\mathcal{A}$ is f and that $\mathcal{A}$ is partitioned into k regions. Then for every k-partition and for every $\varepsilon > 0$, there exists a "3-edge" partition whose within cluster sum of squares differs from that of the given partition by at most ε.

Proof.

Since the within cluster sum of squares, WSS, can be expressed in the form:

$$WSS = \sum_{1}^{k} WSS_i = f \int_{\mathcal{A}_i} r_i^2 d\,(\text{Area}) ,$$

where r_i = distance to ith cluster centroid and $\mathcal{A}_i$ is the ith cluster, it is clear that WSS is a continuous function of the vertices of the k-partition, for every fixed k.

The lemma follows.

Theorem 1:

Let $\mathcal{A}$ be a polygon with a connected interior in $\mathbb{R}^2$.

Suppose that $\mathcal{A}$ has area A and that f is the constant density over $\mathcal{A}$. Let WSS be the minimum within cluster sum of squares over all possible k-partition of $\mathcal{A}$. Then

$$\lim_{k\to\infty} WSS/\left(\frac{fA^2}{k} \cdot \frac{5\sqrt{3}}{54}\right) \geq 1 .$$

Proof.

Let A_i be the area of the ith cluster and E_i be the number of edges of the ith cluster.

[I] We will first obtain an expression for $\sum_{1}^{k} E_i$.

Consider the configuration of the optimal k-partition of $\mathcal{A}$.

Using Lemma 2 and by choosing an arbitrarily small ε, it is enough to examine a "3-edge" k-partition.

Let n be the number of vertices of the polygon $\mathcal{A}$.

Then, using a continuity argument, it can be shown that it is enough to consider partitions with $n = n_2$, where n_2 is the number of vertices in the configuration of the given partition associated with exactly two edges. Let n_3 be the number of vertices associated with 3-edges in the configuration. Using some results in graph theory, we have

(1) $2E = 3n_3 + 2n$ where E is the total number of edges.

Moreover, Euler's formula gives

$E + 1 = F + V$, where F is the number of faces (clusters) and V is the number of vertices.

Therefore, from (1), we have

$$\frac{1}{2}(3n_3 + 2n) + 1 = k + (n_3 + n) ,$$

which gives

(2) $n_3 = 2(k - 1)$.

Hence from (1) and (2),

(3) $$\sum_{1}^{n} E_i = 2E - \text{number of edges around the perimeter}$$

$$= 2E - (n_B + n)$$

$$= 6(k-1) + n - n_B ,$$

where n_B is the number of "3-edge" vertices on the boundary of $\mathcal{A}$.

(Relationship (3) holds for all partitions in which the vertices of the polygon $\mathcal{A}$ have two edges meeting them, and all remaining vertices have three.)

[II] Next, we will find a lower bound for WSS .

Let WSS_i be the within polygon sum of squares of the ith cluster.

Then $WSS = \sum_{1}^{k} WSS_i$.

By Lemma 1, we have,

$$WSS_i \geq \frac{1}{2E_i} f A_i^2 \left[\frac{1}{\tan \Pi/E_i} + \frac{1}{3} \tan \frac{\Pi}{E_i}\right] \quad \text{for all } 1 \leq i \leq k .$$

Therefore,

$$WSS \geq \frac{1}{2} f \sum_{1}^{k} A_i^2 \frac{1}{E_i}\left[\frac{1}{\tan \Pi/E_i} + \frac{1}{3} \tan \frac{\Pi}{E_i}\right] = \frac{1}{2} f \sum_{1}^{k} A_i^2 \, g(E_i) ,$$

where $g(E_i) = \frac{1}{E_i}\left[\frac{1}{\tan \Pi/E_i} + \frac{1}{3} \tan \frac{\Pi}{E_i}\right]$ for $i=1, 2, \ldots, k$.

Now the minimum of $\sum_{1}^{k} A_i^2 g(E_i)$ with all the E_i's being real numbers is not greater than its minimum with all the E_i's being integers, when both are subjected to the constraints:

(i) $\sum_{1}^{k} A_i = A$ and (ii) $\sum_{1}^{k} E_i = 6(k-1) + n - n_B$.

Consider the minimum value of $\sum_{1}^{k} A_i^2 g(E_i)$ with all the E_i's being real numbers. Using Lagrange multipliers, this minimum must satisfy

(iii) $A_i g(E_i) = c_1$, and (iv) $A_i^2 g^{(1)}(E_i) = c_2$,

where c_1 and c_2 are constants independent of i . It follows from (iii) and (iv) that

$$\frac{g^{(1)}(E_i)}{g(E_i)^2} = \frac{c_2}{c_1^2} = \text{constant} \quad \text{for } i=1, \ldots, k .$$

But it can be shown that $\dfrac{g^{(1)}(E_i)}{g(E_i)^2}$, the first derivative of $-1/g$ at E_i, is a monotone decreasing function of E_i.

Therefore, the minimum of $\sum_1^k A_i^2 g(E_i)$ must have

(v) $E_i = \sum_1^k E_i/k = 6 - (n_B + 6 - n)/k$ and (vi) $A_i = A/k$, for all $1 \le i \le k$.

Thus,

$$(4) \qquad WSS = \sum_1^k WSS_i \ge \frac{1}{2k} fA^2 \cdot g(6 - (n_B + 6 - n)/k) .$$

Now, for $k \ge 4$, a lower bound of n_B is 3.

It follows that $6 - (n_B + 6 - n)/k \le 6(1 + (n - 9)/6k)$.

Since $g(E_i)$ is a monotone decreasing function of E_i,

$$g(6 - (n_B + 6 - n)/k) \ge g(6(1 + (n - 9)/6k)) .$$

Therefore, from (4), we have

$$(5) \qquad WSS \ge \frac{1}{2k} fA^2 \cdot g(6(1 + (n - 9)/6k)) .$$

Now for fixed n, $6(1 + (n - 9)/6k) \to 6$ as $k \to \infty$.

Therefore, since g is continuous,

$$\liminf_{k\to\infty} WSS/(\frac{5\sqrt{3}}{54} \cdot \frac{fA^2}{54}) \ge 1 ,$$

and the theorem follows.

Corollary

Let $\mathscr{A}$ be a region with a connected interior in $\mathbb{R}^2$ whose boundary is of finite length. Let A be the area of $\mathscr{A}$ and let $1/A$ be the constant density over $\mathscr{A}$. Let WSS be the minimum within cluster sum of squares over all k partitions of $\mathscr{A}$. Then

$$\lim_{k\to\infty} \mathrm{WSS}/(\frac{A}{k} \cdot \frac{5\sqrt{3}}{54}) \geq 1 .$$

Proof.

For each n, let $\mathscr{A}_n$ be an n-sided polygon of area A_n approximating $\mathscr{A}$ from the inside. Then $A_n/A = 1 + C_n$, where $C_n \to 0$ as $n \to \infty$. Denote the minimum within cluster sum of squares over all k partitions of $\mathscr{A}_n$ by WSS_n. Then, since $\mathscr{A}_n \subset \mathscr{A}$, $\mathrm{WSS}_n \leq \mathrm{WSS}$.

Thus, by putting $f = 1/A$, we have

$$\lim_{k\to\infty} \mathrm{WSS}/(\frac{fA^2}{k} \cdot \frac{5\sqrt{3}}{54}) \geq \lim_{k\to\infty} \mathrm{WSS}_n/(\frac{fA^2}{k} \cdot \frac{5\sqrt{3}}{54})$$

$$= \lim_{k\to\infty} \mathrm{WSS}_n/(\frac{fA_n^2}{k} \cdot \frac{5\sqrt{3}}{54})(1 + C_n) .$$

Letting $n \to \infty$, and using Theorem 1, we obtain

$$\lim_{k\to\infty} \mathrm{WSS}/(\frac{fA^2}{k} \cdot \frac{5\sqrt{3}}{54}) \geq 1 .$$

Theorem 2

Under the hypothesis stated in the Corollary

$$\lim_{k\to\infty} \mathrm{WSS}/(\frac{A}{k} \cdot \frac{5\sqrt{3}}{54}) = 1 .$$

Proof.

Using the Corollary, it is sufficient to show

$$\lim \mathrm{WSS}/(\frac{A}{k} \cdot \frac{5\sqrt{3}}{54} \leq 1 .$$

Now given the region $\mathscr{A}$, we can always construct a region $\mathscr{B}$ of area B consisting

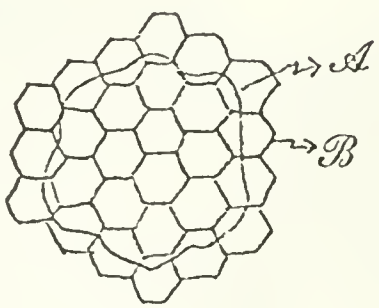


Fig. 5

of k connected regular hexagons (see Fig. 5) such that

(1) $\mathcal{B} \supset \mathcal{A}$, and

(2) $\lim_{k\to\infty} B/A = 1$.

Let $WSS_{\mathcal{B}}$ be the minimum within cluster sum of squares over all k-partitions of $\mathcal{B}$, and let $1/A$ be the constant density over $\mathcal{B}$. Then $\frac{5\sqrt{3}}{54} \cdot \frac{B^2}{Ak} \geq WSS_{\mathcal{B}}$, since the k regular hexagons form a k-partition of $\mathcal{B}$. Now, from (1), $WSS_{\mathcal{B}} \geq WSS$, and hence

$$\frac{5\sqrt{3}}{54} \cdot \frac{B^2}{Ak} \geq WSS .$$

Thus,

(3) $WSS/(\frac{5\sqrt{3}}{54} \cdot \frac{A}{k}) \leq B^2/A^2$,

and the theorem follows from (2) and (3).

The result of Theorem 2 only gives a lower bound for the overall within cluster sum of squares of the k-means partition. It falls short in showing that the within sum of squares of the k clusters are asymptotically equal (a conjecture due to Professor John A. Hartigan). Much work has yet to be done to prove the conjecture for two or more dimensional distributions.

3. EMPIRICAL ILLUSTRATIONS

In order to illustrate the asymptotic properties of bivariate k-means clusters obtained in Section 2, an empirical study is performed using bivariate samples generated according to the uniform distribution on the unit square. It is necessary to estimate the within cluster sum of squares WSS for various values of k from generated samples because the WSS for the optimal k-mean partition of the unit square cannot be obtained analytically for large values of k. Here, the results of three sets of experiments using different sample sizes are reported.

In Experiment One, four different samples of size $N = 1500$ are generated from the uniform distribution on the unit square. Using $k = 40$, 50, 60, and 70 for the different samples, unbiased estimates WSS_N of WSS for the different cluster sizes are obtained. The values of WSS_N for the various values of k are given alongside the corresponding asymptotic lower bounds for WSS (that is, $\frac{5\sqrt{3}}{54} \cdot \frac{1}{k}$) in Table 1, and the corresponding pairs are found to be in close agreement with one another. Similarly, three different samples of size $N = 2500$ are generated in Experiment Two, and the values of k used for the three samples are $k = 50$, 60, and 70; while in Experiment Three, the three generated samples are of size $N = 4000$, and the values of k used are $k = 60$, 80, and 100. The resulting WSS_N values for these six experimental trials are also given in Table 1. Again, the values of WSS_N for the various values of k are found to be in close agreement with the corresponding lower bounds for WSS. Hence, these empirical results tend to indicate that the asymptotic lower bound obtained in Theorom 2 is the WSS for the optimal k-means partition when k becomes large.

Table 1

Sample Size (N)	k	$WSS_N (\times 10^{-3})$	$\frac{5\sqrt{3}}{54} \cdot \frac{1}{k} \ (\times 10^{-3})$
1500	40	4.059	4.009
1500	50	3.208	3.208
1500	60	2.686	2.673
1500	70	2.259	2.291
2500	50	3.253	3.208
2500	60	2.733	2.673
2500	70	2.393	2.291
4000	60	2.659	2.673
4000	80	2.035	2.005
4000	100	1.611	1.604

The graph configurations of two of the sample k-means partitions are given in Figures 1 and 2. In Figure 1, the sample size used is 1500 and $k = 50$, while in Figure 2, the sample size used is 4000 and $k = 100$. Although only a few regular hexagons appear in these two configurations, it seems plausible that, when k is large enough, the graph configuration of optimal k-means partition would consist mostly of regular hexagons.

ACKNOWLEDGEMENTS

This research was supported in part by the National Science Foundation, Grant No. MCS75-08374. The author would like to thank John A. Hartigan and David Pollard for many useful discussions.

BIBLIOGRAPHY

Hartigan, J.A. (1975). Clustering Algorithms. New York: John Wiley.

Hartigan, J.A. (1978). Asymptotic distributions for clustering criteria. Annals of Statistics, 6, 117-131.

Hartigan, J.A. and Wong, M.A. (1979a). Hybrid Clustering. Proceedings of the 12th Interface Symposium on Computer Science and Statistics, ed. Jane Gentleman, University of Waterloo Press, 137-143.

Hartigan, J.A. and Wong, M.A. (1979b). Algorithm AS 136: A K-means clustering algorithm. Applied Statistics, 28, 100-108.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings: Fifth Berkeley Symposium on Math. Statist. Prob. 1, 281-297.

Pollard, D. (1981). Strong consistency of k-means clustering. Annals of Statistics, 9, 135-140.

Wong, M.A. (1980). Asymptotic Properties of k-means clustering algorithm as a density estimation procedure. Sloan School of Management Working Paper #2000-80, M.I.T., Cambridge.

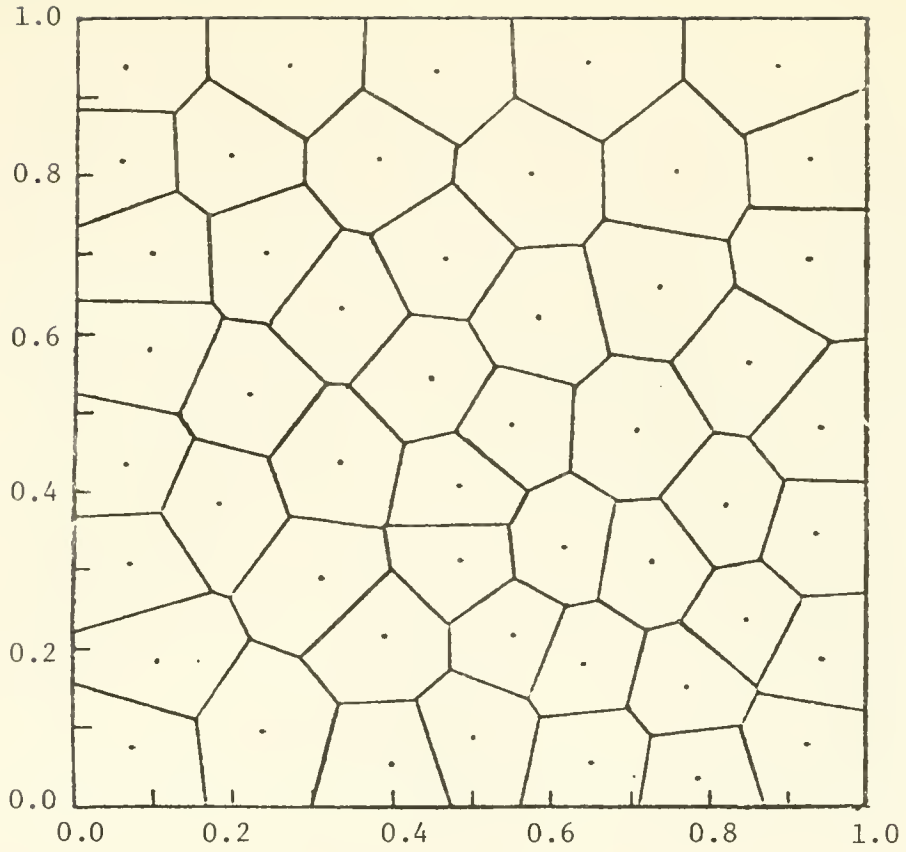


Figure 1. Graph Configuration of sample k-means partition (k = 50) obtained for 1500 observations from the uniform distribution.

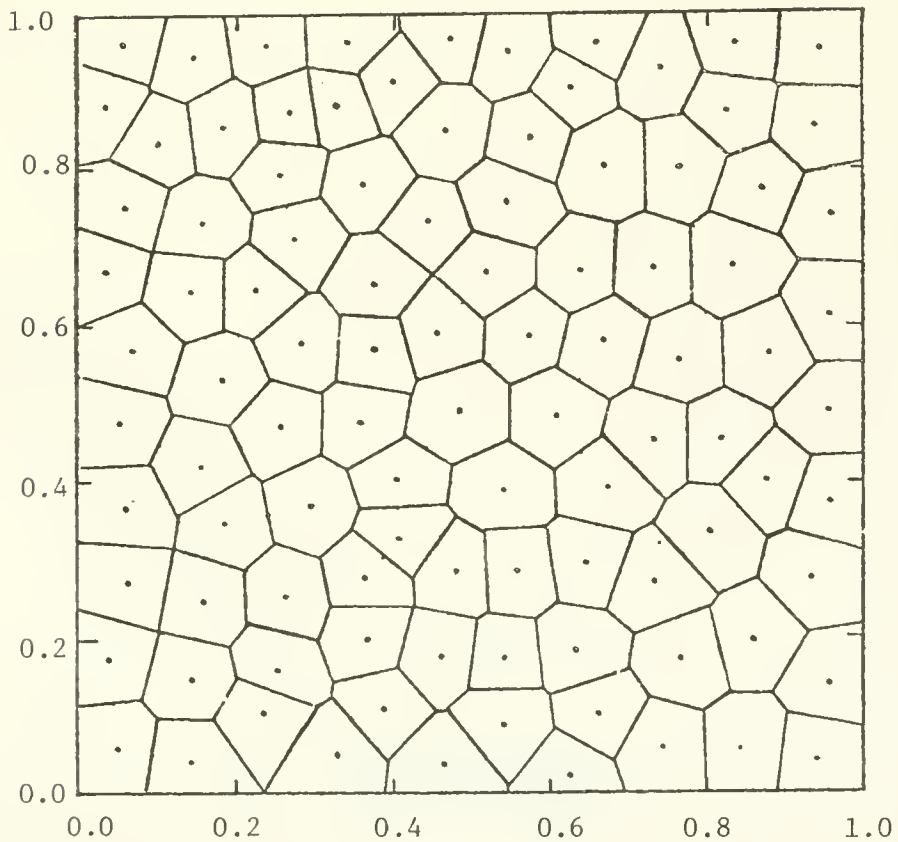


Figure 2. Graph configuration of sample k-means partition (k = 100) obtained for 4000 observations from the uniform distribution.